ПСМО ФКН ВШЭ, 3 курс, 1 модуль

Задание 1. Основы статистики и бутстрап.

Прикладная статистика в машинном обучении, осень 2018

Время выдачи задания: 20 сентября (четверг).

Срок сдачи: **4 октября (четверг), 23:59.**

Среда для выполнения практического задания – PYTHON 2.x/PYTHON 3.x.

Правила сдачи

Инструкция по отправке:

1. Решения задач следует присылать единым файлом формата `pdf`, набранным в LaTeX, либо в составе ipython-тетрадки в форматах `ipynb` и `html` (присылайте оба формата, т.к. AnyTask из-за высокой загрузки иногда не рендерит тетрадки в формате `ipynb` – а если мы не увидим ваши задачи, мы их не проверим). Отправляйте практические задачи в виде отдельных файлов (ipython-тетрадок или исходных файлов с кодом на языке `python`).

Оценивание и штрафы:

1. Максимально допустимая оценка за работу над основными задачами – 10 баллов.

2. Бонусные баллы (см. конец домашнего задания) и влияют на освобождение от задач на экзамене.

3. Дедлайн жесткий. Сдавать задание после указанного срока сдачи нельзя.

4. Задание выполняется каждым студентом индивидуально и независимо от других студентов. «Похожие» решения считаются плагиатом и все студенты (в том числе те, у кого списали) не могут получить за него больше 0 баллов, причем обнуляются и бонусные баллы. Если вы нашли решение какого-то из заданий (или его часть) в открытом источнике, необходимо указать ссылку на этот источник в отдельном блоке в конце вашей работы (скорее всего вы будете не единственным, кто это нашел, поэтому чтобы исключить подозрение в плагиате, необходима ссылка на источник).

Основные задачи

1. (1 балл) Дано множество из n элементов. Написать программу, печатающую на экране случайное подмножество этого множества, состоящее из k элементов, причем все C_n^k подмножеств равновероятны.

2. (1 балл) Подсчитать математическое ожидание числа сравнений при сортировке n различных чисел алгоритмом QuickSort, если в исходном массиве числа находятся в случайном порядке.

3. (3 балла) Пусть есть выборка из 11 элементов: $x_{(1)} < x_{(2)} < x_{(3)} < x_{(4)} < x_{(5)} < x_{(6)} < x_{(7)} < x_{(8)} < x_{(9)} < x_{(10)} < x_{(11)}$. Оцениваемая статистика θ – медиана.

   (a) (1 балл) Покажите что для оценки $\widehat{\theta}$ по бутстрепной выборке верно следующее:

   $$\mathrm{P}(\widehat{\theta} > x_{(i)}) = \sum_{j=0}^{5} \mathrm{Bin}\left(j, n, \frac{i}{n}\right),$$

   где $\mathrm{Bin}(j; n, p) = C_n^j p^j (1-p)^{n-j}$.

   (b) (1 балл) Покажите, что оценка $\widehat{\theta}$ по бутстрепной выборке равна $x_{(i)}$ с вероятностью:

   $$P\left(\widehat{\theta} = x_{(i)}\right) = \sum_{j=0}^{5} \left(\mathrm{Bin}\left(j; n, \frac{i-1}{n}\right) - \mathrm{Bin}\left(j, n, \frac{i}{n}\right)\right),$$

   (c) (1 балл) Используя результат пункта 3a, подсчитайте 90% бутстрепный доверительный интервал для медианы. *Подсказка:* посчитайте $\mathrm{P}(\widehat{\theta} \leqslant 3)$ и $\mathrm{P}(\widehat{\theta} \geqslant 9)$.

4. (2 балла) Скачайте данные по ссылке https://vincentarelbundock.github.io/Rdatasets/csv/MASS/galaxies.csv. Данные состоят из скоростей 82 галактик из созвездия Северной Короны. Мы хотим узнать, есть ли пустоты или суперкластеры в данной части вселенной. Одним из свидетельств наличия пустот и кластеров в данных является многомодальность распределения скоростей галактик. Другими словами, нам необходимо проверить гипотезу унимодальности распределения, т.е.:

$$H_0 : n_{\text{mode}}(p) = 1 \qquad \text{против альтернативы} \qquad H_a : n_{\text{mode}}(p) > 1.$$

Плотность распределения будем оценивать напараметрическим ядерным методом:

$$\widehat{p}_{K,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

где $K(x) = \exp\{-x^2\}$ – гауссово ядро, причем h – ширина этого ядра. Таким образом, $K\left(\frac{x-X_i}{h}\right)$ – это ядро ширины h, «помещенное» на точку выборки X_i.

(a) (1 балл) По данным найдите минимальное $\widehat{h}_{\text{uni}}$, при котором распределение ещё унимодально. Найденная $\widehat{h}_{\text{uni}}$ является оценкой по данным для реальной h_{uni}. Если окажется, что $h_{uni} > \widehat{h}_{\text{uni}}$, то это значит, что в реальности мод больше одной. Т.е. нулевая гипотеза отвергается на уровне значимости α:

$$\mathrm{P}(\text{multimodal}) = \mathrm{P}(h_{\text{uni}} > \widehat{h}_{\text{uni}}) \leqslant \alpha. \tag{1}$$

(b) (1 балл) Используя бутстреп, оцените следующую величину:

$$\widehat{\mathrm{P}}(h_{\text{uni}} > \widehat{h}_{\text{uni}}) \approx \frac{1}{B} \sum_{b=1}^{B} \left(\widehat{h}^{b}_{\text{uni}} \geqslant \widehat{h}_{\text{uni}}\right) \tag{2}$$

Сэмплирование будем делать из непараметрической ядерной смеси. Для этого обоснуйте, что сэмплирование из смеси эквивалентно следующей схеме сэмплирования: $X_j^* \sim X_i + \widehat{h}_{\text{uni}}\mathcal{N}(0,1)$, где X_i – случайный элемент оригинальной выборки.

N.B. так как сэмплирование делается не из оригинальной эмпирической выборки, а из сглаженной, то дисперсия стала выше. Как нужно скорректировать предложенную схему сэмплирования, чтобы дисперсия не изменилась?

Подсказка: для схемы сэмплирования $X_j^* \sim a + b(X_i + \widehat{h}_{\text{uni}}\mathcal{N}(0,1))$ найдите такие a и b, что первый и второй моменты этого распределения и эмпирического распределения совпадают.

5. (3 балла) Скачайте данные по ссылке https://vincentarelbundock.github.io/Rdatasets/csv/boot/cd4.csv.

   Датасет CD4 содержит информацию о 20 ВИЧ-инфицированных пациентах до и после года лечения на экспериментальном антивирусном лекарстве (см. заголовок таблицы).

   (a) (1 балл) Для коэффициента корреляции Пирсона между данными до и после лечения посчитайте 95% доверительный интервал следующими методами: нормальный, перцентильный, центральный и t-бутстреп. Для подсчёта дисперсии $\widehat{\sigma}$ для t-bootstrap используйте следующую формулу: $\widehat{\sigma}(r) = \sqrt{\frac{1-r^2}{n-2}}$.

   Очень часто для работы бывает удобно применить нормализующее преобразование: $\frac{1}{2}\log\frac{1+r}{1-r}$. Сделайте это и посчитайте заново все интервалы. Что изменилось? Стали ли они более согласованными? Для подсчёта дисперсии для t-bootstrap в нормализованном виде используйте следующую оценку дисперсии: $\widehat{\sigma}(r) = \frac{1}{\sqrt{n-3}}$.

(b) (1 балл) В предыдущем пункте для t-bootstrap вы использовали некоторое аналитическое приближение для вычисления дисперсии корреляции. Теперь вам предстоит реализовать двойной бутстрап: над бутстрепными выборками по которым считается r делайте ещё бутстреп для оценки $\sigma(r)$. Сравните дисперсии, подсчитанные аналитически и бутстрепом. Что вы заметили? Как изменились доверительные интервалы?

(c) (1 балл) После предыдущего пункта вам наверняка захотелось проверить все наши оценки на смещённость. Оцените смещение (bias) для коэффициента корреляции с помощью jackknife.

Бонусные задачи

1. (1 бонусный балл) Пусть $T_n = \overline{X}_n^2$, $\mu = \mathrm{E}\,(X_1)$, $\alpha_k = \int |x-\mu|^k dF(x)$ и $\widehat{\alpha}_k = \sum\limits_{i=1}^{n} \left|X_i - \overline{X}_n\right|^k$. Докажите, что оценка дисперсии функционала T_n с помощью бутстрепа равна:

$$v_{boot} = \frac{4\overline{X}_n^2 \widehat{\alpha}_2}{n} + \frac{4\overline{X}_n \widehat{\alpha}_3}{n^2} + \frac{\widehat{\alpha}_4}{n^3} + \frac{\widehat{\alpha}_2^2(2n-3)}{n^3}$$

2. (1 бонусный балл) Доказать эффективность бэггинга можно на следующем игрушечном примере. Рассмотрим задачу классификации и предиктор («решающий пень», т.е. решающее дерево глубины 1) вида:

$$\widehat{\theta}_n(x) = \mathbf{1}_{[\widehat{d}_n \leqslant x]}, \quad x \in \mathbb{R}.$$

Здесь $\widehat{d}_n$ – действительное число, оцененненное по выборке $\mathbf{X}^\ell = \{Y_i, X_i\}_{i=1}^{\ell}$. Пусть оценка $\widehat{d}_n$ асимптотически нормальна, причем скорость сходимости к нормальному распределению у нее b_n^{-1}, т.е.

$$b_n(\widehat{d}_n - d_0) \to_D \mathcal{N}(0, \sigma_\infty^2),$$

где σ_∞^2 – ее асимптотическая дисперсия.

Рассмотрим некоторый x в b_n^{-1}-окрестности параметра d_0, т.е. $x = x_n(c) = d_0 + c\sigma_\infty b_n^{-1}$.

Подсчитайте: (1) асимптотические математическое ожидание и дисперсию классификатора $\widehat{\theta}_n(x)$ для таких x; (2) асимптотические математическое ожидание и дисперсию бэггинг-классификатора $\widehat{\theta}_{n;B}(x) = \frac{1}{J}\sum\limits_{j=1}^{J} \widehat{\theta}_{n;(j)}(x)$ для таких x. Асимптотики рассматривать при $n \to \infty$. Что можно сказать, сравнивая дисперсии обычного и бэггинг-классификатора?